Pathology-based Ordinal Multiple Instance Learning Framework for Esophageal Cancer Tumor Regression Grading

Lin Zhao[1†], Zhengjin Liu[2†], Weiping Lin[1], Baptiste Magnier[3, 4], Ziqing Zhao[2], Jun Zhong[5], Yong Wang[5], Shouguo Li[6], Haijie Lu[6*], Liansheng Wang[1*]

[1]Department of Computer Science, School of Informatics, Xiamen University, Xiamen, China.
[2]Department of Pathology, Zhongshan Hospital of Xiamen University, School of Medicine, Xiamen University Xiamen, China.
[3]EuroMov Digital Health in Motion, Univ Montpellier, IMT Mines Ales, Ales, France.
[4]Service de Médecine Nucléaire, Centre Hospitalier Universitaire de Nîmes, Université de Montpellier, Nîmes, France.
[5]Department of Radiology, Zhongshan Hospital of Xiamen University, School of Medicine, Xiamen University Xiamen, China.
[6]Department of Radiation Oncology, Zhongshan Hospital of Xiamen University, School of Medicine, Xiamen University Xiamen, China.

*Corresponding author(s). E-mail(s): luhaijie@xmu.edu.cn; lswang@xmu.edu.cn;
†These authors contributed equally to this work.

Abstract

Tumor Regression Grade (TRG) is a critical pathological indicator for evaluating treatment response and guiding subsequent therapy in esophageal cancer. It fundamentally relies on the descriptive assessment of the reciprocal progression between residual tumor and regression traces like fibrosis. However, current TRG assessment relies on subjective interpretation of H&E-stained slides, which is labor-intensive and prone to inter-observer variability. Although Multiple Instance Learning (MIL) has been widely adopted for whole-slide image analysis, existing methods rarely account for the specific pathological characteristics

of TRG, often neglecting critical regressive signs and the intrinsic ordinality of the grading process. To address these limitations, we propose a novel MIL framework for esophageal cancer TRG assessment. We introduce a Dual-Branch Gated Attention architecture to disentangle malignant cellular patterns from regressive traces, emulating the diagnostic rationale of pathologists who contrast tumor residues against regression signs during assessment. Additionally, an Ordinal-Aware Regression module is incorporated to regularize the feature space to reflect ordinal continuity. Extensive experiments show that our framework significantly outperforms state-of-the-art MIL methods, highlighting its effectiveness. Our code will be made publicly available soon.

Keywords: Tumor Regression Grade, Multiple Instance Learning, Esophageal Cancer.

1 Introduction

Esophageal cancer is the seventh leading cause of cancer-related mortality worldwide [1]. Neoadjuvant chemoradiotherapy (nCRT) followed by surgery is widely adopted in clinical practice, as it effectively reduces tumor burden and downstages disease [2, 3]. However, highly heterogeneity exists in pathological response to nCRT. To standardize response evaluation, tumor regression grading (TRG) was established as a histopathological scoring system [4]. TRG is a strong prognostic indicator, making its accurate assessment critical for guiding postoperative management and improving survival stratification [5].

According to the eighth edition of the American Joint Committee on Cancer (AJCC) staging system [6], TRG fundamentally relies on the descriptive evaluation of the relationship between residual tumor and treatment-induced regressive traces (e.g., fibrosis). Current TRG evaluation for esophageal cancer mainly relies on pathologists' subjective assessment of H&E-stained slides, which is labor-intensive and suffers from inter-observer variability [7]. To address this, deep learning has explored automated TRG modeling through cell-level graph neural networks [8] to capture cellular interactions. However, TRG is inherently a pathology-based grading process that requires holistic assessment of the interplay between residual tumor and regressive traces across the entire slide. Cell-level models may overemphasize local features and require fine-grained annotations that are difficult to obtain.

Weakly-supervised Multiple Instance Learning (MIL) on whole-slide images (WSIs) has emerged as an effective framework for capturing the full morphological spectrum using only slide-level labels [9–12]. Representative methods include attention-based models such as ABMIL [9] and DeepAttnMISL [13], the clustering-enhanced CLAM [10], and Transformer-based architectures like TransMIL [14] and HIPT [15], which are designed to model global and hierarchical dependencies. However, this shift to MIL often comes at the cost of modeling fidelity. By operating at the slide level, generic MIL models struggle to capture the fine-grained morphological patterns that dedicated, fully-supervised frameworks could leverage, making them suboptimal for nuanced tasks like TRG assessment. Specifically, most MIL models emphasize tumor

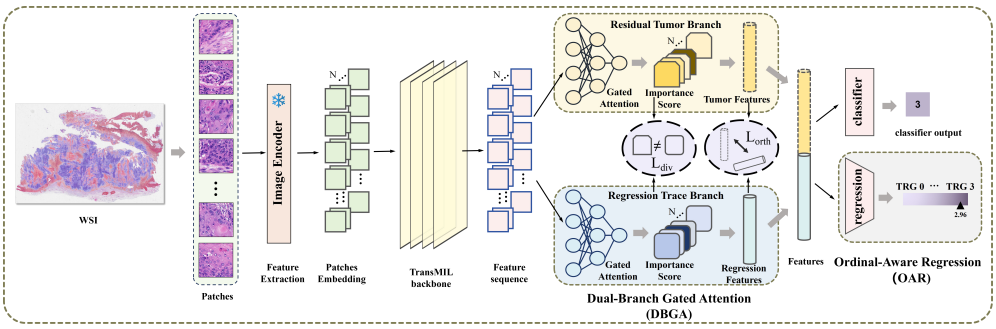

Fig. 1: Overview of our framework. The DBGA utilizes two branches to independently capture disentangled representations of malignant cells and regression traces, subsequently modeling their interaction. The OAR further regularizes the feature space to preserve the intrinsic ordinality of TRG.

regions while neglecting regressive signs like fibrosis, which provide critical evidence for TRG assessment. Furthermore, these methods treat TRG as independent classes, ignoring its intrinsic ordinality and leading to prediction errors that violate clinical logic.

We propose a novel MIL framework for esophageal cancer TRG assessment (Fig. 1). To capture pathology-based diagnostic knowledge regarding the reciprocal progression between residual tumor and regression signs, we design a Dual-Branch Gated Attention (DBGA) architecture, comprising a Residual Tumor Branch and a Regressive Trace Branch. Guided by orthogonality and diversity constraints, DBGA disentangles malignant cellular patterns from regression traces. This design emulates pathologists' diagnostic reasoning by focusing on the contrast between tumor residues and regression traces. Furthermore, to account for the ordinality among TRG, we introduce an Ordinal-Aware Regression (OAR) module. By incorporating an auxiliary regression task into the primary classification task, OAR enforces ordinal consistency in feature representations, mitigating clinically implausible label jumps (e.g., predicting TRG 1 from TRG 3) and improves overall assessment reliability. Our contributions are as follows: 1) We propose a novel MIL framework to address the neglect of regression signs and the inherent ordinality of TRG in esophageal cancer assessment. 2) We introduce DBGA to disentangle malignant cellular patterns from tissue regression traces, and OAR to preserve intrinsic ordinality of TRG. 3) Extensive experiments show that our framework effectively captures the interplay between tumor residues and regression traces, which substantially improves the accuracy and reliability of TRG assessment.

2 Methodology

2.1 Preliminaries and Feature Encoding

Following the standard MIL paradigm, each WSI is represented as a bag $X = \{x_1, \ldots, x_N\}$ with a slide-level label $Y \in \{0, 1, 2, 3\}$ (corresponding to TRG grades 0–3), where $x_i \in \mathbb{R}^d$ denotes the d-dimensional feature vector of the i-th patch. To

model tissue heterogeneity, we adopt TransMIL [14] as the backbone to capture both local morphology and global dependencies. Specifically, spatial context is incorporated into the feature bag X using a Pyramid Position Encoding Generator, which captures multi-scale positional information. The resulting sequence is then processed by N_{layer} Transformer layers to produce context-aware patch features $H \in \mathbb{R}^{N \times d}$. These features serve as input to the subsequent dual-branch attention module.

2.2 Dual-Branch Gated Attention Architecture

To capture the reciprocal progression between residual tumor tissue and treatment-induced regressive traces, we designed the DBGA architecture. This design emulates the diagnostic rationale used by pathologists during TRG assessment by contrasting residual malignant features against reactive stromal changes. DBGA consists of two parallel branches: a **Residual Tumor Branch** dedicated to malignant cellular patterns, and a **Regression Trace Branch** dedicated to various signs of tissue regression such as fibrosis. For each branch $k \in \{\text{tumor, regression}\}$, the attention score $a_{k,i}$ for patch feature h_i is computed using a gated attention mechanism [9] commonly adopted in MIL:

$$a_{k,i} = \frac{\exp\left\{w_k^\top \left(\tanh(V_k h_i^\top) \odot \sigma(U_k h_i^\top)\right)\right\}}{\sum_{j=1}^{N} \exp\left\{w_k^\top \left(\tanh(V_k h_j^\top) \odot \sigma(U_k h_j^\top)\right)\right\}}, \tag{1}$$

where $V_k, U_k \in \mathbb{R}^{L \times d}$ and $w_k \in \mathbb{R}^L$ are learnable parameters. The attention scores are used to aggregate patch features into a branch-specific representation: $Z_k = \sum_{i=1}^{N} a_{k,i} h_i \in \mathbb{R}^d$. To capture the reciprocal relationship between tumor residues and regressive traces emphasized in AJCC, these branch-specific representations are integrated into a comprehensive slide-level embedding:

$$Z_{bag} = \text{Concat}(Z_{tumor}, Z_{regression}) \in \mathbb{R}^{2d}. \tag{2}$$

Given only slide-level labels, we introduce structured priors to prevent the model from learning an arbitrary and biologically meaningless feature separation. Since TRG is fundamentally based on two morphologically distinct components—residual tumor and regressive traces—we design a feature disentanglement mechanism in the latent space. This is achieved through an orthogonality loss, which encourages the two branches to capture independent information, and a spatial diversity loss, which promotes attention on different regions.

Orthogonality Constraint. To ensure that the two branches capture non-overlapping semantic information, the orthogonality constraint encourages the model to learn disentangled representations. This is formalized as the squared cosine similarity between Z_{tumor} and $Z_{regression}$:

$$\mathcal{L}_{orth} = \left(\frac{Z_{tumor}^\top Z_{regression}}{||Z_{tumor}||_2 \cdot ||Z_{regression}||_2} \right)^2. \tag{3}$$

Minimizing $\mathcal{L}_{orth}$ encourages the model to learn representations that are semantically aligned with the clinical definition of TRG. Consequently, it steers one branch to focus

on malignant cellular patterns while the other focuses on regressive traces, particularly fibrosis.

Diversity Constraint. To prevent the two branches from focusing on the same spatial regions within the WSI, we introduce a diversity loss based on the attention maps. Let $A_k = [a_{k,1}, a_{k,2}, \ldots, a_{k,N}]^\top \in \mathbb{R}^N$ denote the attention weight vector for the k-th branch. We encourage the attention distributions A_{tumor} and $A_{regression}$ to be dissimilar by minimizing their cosine similarity:

$$\mathcal{L}_{div} = \frac{A_{tumor}^\top A_{regression}}{||A_{tumor}||_2 \cdot ||A_{regression}||_2}. \tag{4}$$

Minimizing $\mathcal{L}_{div}$ ensures spatial separation, encouraging the branches to focus on distinct malignant and regressive patterns.

2.3 Ordinal-Aware Regression Module

The morphological difference between TRG 0 and TRG 3 is substantially greater than that between adjacent grades, reflecting the inherent ordinality of TRG. However, standard classification heads are typically designed to learn discriminative boundaries by mapping features to distinct class prototypes, aiming for maximum separability between categories. This orthogonalization of categories often leads the model to ignore the continuous biological spectrum underlying TRG, potentially resulting in ordinality-violating prediction errors.

To address this limitation, we introduce OAR. Alongside the classification head, the fused embedding Z_{bag} is fed into a regression head to predict a continuous score $\hat{y}_{reg} \in \mathbb{R}$. Unlike the classification head, which optimizes for categorical separation, OAR regularizes the feature space to reflect ordinal continuity. By minimizing the Mean Squared Error, the model encourages samples with similar TRG grades to be mapped to nearby regions in the feature space while preserving the continuum from favorable to poor treatment response:

$$\mathcal{L}_{reg} = \frac{1}{N_{batch}} \sum_{k=1}^{N_{batch}} (Y^{(k)} - \hat{y}_{reg}^{(k)})^2, \tag{5}$$

where N_{batch} denotes the number of samples in a training batch, and $\hat{y}_{reg}^{(k)}$ represents the predicted continuous score for the k-th slide. This objective provides an ordinal smoothness prior, encouraging predictions to respect the disease progression of TRG and reducing implausible cross-grade jumps.

Optimization Objective. The final training objective is formulated as:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \lambda_{reg}\mathcal{L}_{reg} + \lambda_{orth}\mathcal{L}_{orth} + \lambda_{div}\mathcal{L}_{div}, \tag{6}$$

where $\mathcal{L}_{cls}$ denotes the cross-entropy loss, $\lambda_{reg}, \lambda_{orth}$, and λ_{div} are hyperparameters. By minimizing $\mathcal{L}_{total}$, the model learns to capture the interplay between malignant cellular patterns and regression traces. Simultaneously, it preserves the ordinality

Table 1: Performance comparison with ten MIL methods for TRG assessment. Results are reported as mean ± standard deviation across 5-fold cross-validation. Statistical comparisons between methods are performed using Wilcoxon signed-rank tests, with $p < 0.05$ considered significant.

Method	Accuracy	F1	AUC	p-value
Max-pooling [16]	$0.5417_{\pm 0.0589}$	$0.5301_{\pm 0.0738}$	$0.7606_{\pm 0.0540}$	< 0.01
Mean-pooling [16]	$0.5083_{\pm 0.0543}$	$0.4585_{\pm 0.0609}$	$0.7597_{\pm 0.0497}$	< 0.01
ABMIL [9]	$0.6250_{\pm 0.1531}$	$0.6171_{\pm 0.1573}$	$0.8468_{\pm 0.0639}$	< 0.01
CLAM [10]	$0.7000_{\pm 0.0950}$	$0.6964_{\pm 0.0957}$	$0.8995_{\pm 0.0648}$	< 0.01
DeepAttnMISL [13]	$0.6667_{\pm 0.0589}$	$0.6298_{\pm 0.0586}$	$0.8778_{\pm 0.0426}$	< 0.01
TransMIL [14]	$0.7667_{\pm 0.0812}$	$0.7542_{\pm 0.0823}$	$0.9102_{\pm 0.0488}$	< 0.01
ILRA-MIL [17]	$0.5500_{\pm 0.1079}$	$0.5344_{\pm 0.1187}$	$0.7829_{\pm 0.0428}$	< 0.01
HIPT [15]	$0.7833_{\pm 0.0349}$	$0.7762_{\pm 0.0372}$	$0.9143_{\pm 0.0350}$	< 0.01
ACMIL [18]	$0.7250_{\pm 0.0757}$	$0.7239_{\pm 0.0713}$	$0.8963_{\pm 0.0364}$	< 0.01
AEM [19]	$0.6500_{\pm 0.1401}$	$0.6318_{\pm 0.1540}$	$0.8579_{\pm 0.0606}$	< 0.01
Ours	$\mathbf{0.8167}_{\pm 0.0864}$	$\mathbf{0.8098}_{\pm 0.0933}$	$\mathbf{0.9329}_{\pm 0.0492}$	–

of grades, enabling a precise TRG assessment that aligns with clinical diagnostic reasoning.

3 Experiments

3.1 Datasets and Implementation Details

Our internal dataset consists of 120 H&E slides obtained from 58 patients. The external dataset included 74 H&E slides from 39 patients. Each WSI was assigned a TRG by two experienced gastrointestinal pathologists following the eighth edition of the AJCC staging system [6]. These grades served as the ground truth labels: TRG 0 indicates pathological complete response, TRG 1 near-complete response, TRG 2 partial response, and TRG 3 poor response.

For data preprocessing, we followed the pipeline described in CLAM [10], including tissue segmentation and patch extraction at 20× magnification with a 256×256 pixels patch size. For feature extraction, we employed the H-Optimus-1 [20] model to generate patch-level embeddings of dimension 1536. We updated the model weights using the Adam optimizer with a learning rate of 1×10^{-4} and a weight decay of 1×10^{-5}. All experiments were conducted using a patient-wise five-fold cross-validation strategy. We report the mean and standard deviation of the accuracy, F1 score, and AUC averaged over the five folds.

Table 2: Ablation experiments. *w/o DBGA* and *w/o OAR* denote models without DBGA and OAR, respectively.

Method	Accuracy	F1	AUC
w/o DBGA	0.7917	0.7869	0.9306
w/o OAR	0.7833	0.7733	0.9148
Ours	**0.8167**	**0.8098**	**0.9329**

Table 3: Impact of different hyperparameter values on model performance.

Value	λ_{orth}		λ_{div}		λ_{reg}	
	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
0.1	0.8166	0.9162	0.7917	0.9264	0.8000	0.9190
0.3	0.8000	0.9102	**0.8167**	**0.9329**	0.7334	0.9167
0.5	**0.8167**	**0.9329**	0.7833	0.9116	**0.8167**	**0.9329**
0.7	0.7917	0.9227	0.7833	0.9116	0.7833	0.9158
0.9	0.7583	0.9153	0.7833	0.9060	0.7750	0.9180

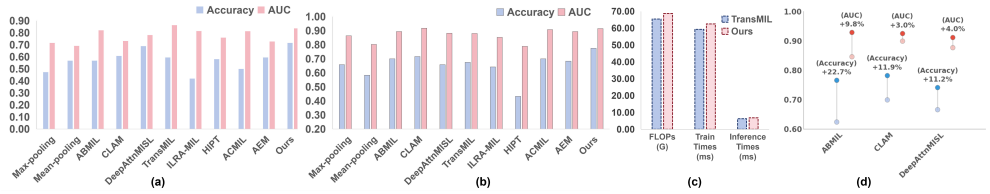

Fig. 2: Generalization and efficiency analysis. (a) Performance on the external cohort. (b) Performance using features extracted by Virchow foundation model. (c) FLOPs and time per slide versus TransMIL. (d) Plug-and-play capability of DBGA and OAR across various MIL architectures. Dark-colored dots represent performance with the DBGA and OAR modules, while light-colored dots indicate the baseline performance.

3.2 Baseline Comparison

We compare our method with eight representative MIL methods, including ABMIL [9], CLAM [10], DeepAttnMISL [13], TransMIL [14], ILRA-MIL [17], HIPT [15], ACMIL [18], and AEM [19], as well as two conventional MIL pooling strategies (Max-pooling and Mean-pooling) [16]. All methods were evaluated under the same experimental setting. Table 1 presents the performance of different methods. Compared to traditional pooling operations, attention-based and Transformer-based models demonstrate superior performance. This stems from their ability to effectively isolate sparse diagnostic signals from extensive background noise. Our method effectively assesses TRG by learning disentangled representations of residual tumor and regression traces, thereby capturing the interplay between malignant cellular patterns and regressive changes. By further incorporating the intrinsic ordinality of regression grades, our framework achieves superior performance across all metrics, yielding a 5% improvement in accuracy over the TransMIL backbone (0.7667 vs. 0.8167). All p-values are below 0.05, confirming that the performance gains of our framework over the competing methods are statistically significant.

3.3 Ablation Study

Effectiveness of DBGA and OAR We conduct ablation studies to examine the contribution of each component. As shown in Table 2, removing DBGA (*w/o DBGA*), implemented by replacing the dual-branch attention with a single attention branch, degrades performance. Removing OAR (*w/o OAR*), which excludes ordinal regularization, also reduces performance. These results confirm the effectiveness of pathology-based feature disentanglement and ordinal supervision for TRG assessment.
Hyperparameter Sensitivity Analysis We analyze the impact of hyperparameters in Table 3. Peak performance occurs at $\lambda_{orth} = 0.5$, as excessive orthogonality distorts the semantic manifold. A small λ_{div} causes overlapping branch attention, whereas a large value shifts attention to irrelevant regions. Optimal performance is achieved at $\lambda_{div} = 0.3$. A low λ_{reg} fails to adequately regularize the feature space according to the ordinal structure of TRG, while excessive weighting oversimplifies decision boundaries.

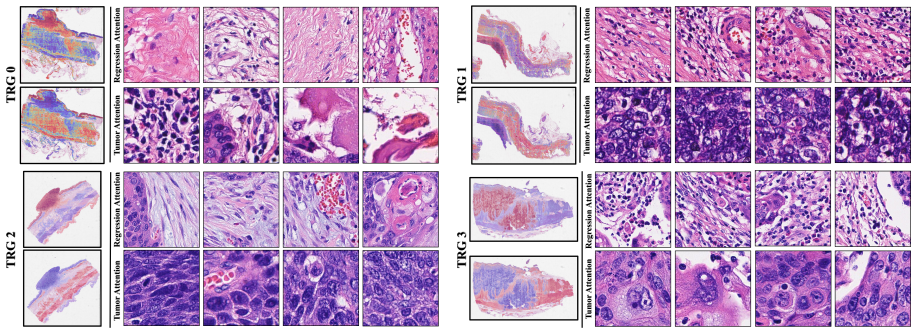

Fig. 3: Qualitative analysis for representative cases across TRG 0–3. Warm colors indicate high attention.

3.4 Generalizability and Computational Efficiency

External Validation We further validated the generalizability of our framework on an independent external cohort. As shown in Fig. 2(a), our method achieves a substantial performance advantage on this external cohort. This demonstrates its excellent generalization to unseen data from different centers, despite variations in tissue preparation and scanning protocols across institutions.

Robustness To assess robustness to different feature extractors, Fig. 2(b) compares methods using the Virchow foundation model [21]. Our framework consistently outperforms others, demonstrating robustness to embedding variations. This indicates that the gains arise from the proposed architecture rather than reliance on a specific extractor, enabling the effective exploitation of diverse semantic embeddings for accurate TRG assessment.

Plug-and-play To validate the plug-and-play capability of DBGA and OAR, we incorporated them into multiple MIL backbones. As shown in Fig. 2(d), consistent performance gains across architectures demonstrate the general applicability of the proposed modules for TRG assessment rather than backbone-specific designs.

Computational Efficiency We evaluate the computational efficiency of our framework in Fig. 2(c). The TransMIL backbone requires 65.37G FLOPs per slide, and introducing DBGA and OAR increases computation by only 5.0%, indicating minimal overhead. Training time rises slightly from 59.32 ms to 62.50 ms per slide, and inference from 6.28 ms to 6.80 ms, maintaining a favorable efficiency–performance trade-off.

3.5 Attention Map Visualization

Fig. 3 illustrates the interpretability of our framework on representative cases. It shows how the model adaptively adjusts attention to reflect changes in tissue composition from TRG 0 to TRG 3. In TRG 0, the tumor branch responds primarily to acellular keratin and calcification, while the regression branch highlights vascular proliferation and multinucleated giant cells, confirming complete regression. In TRG 1–2, the tumor branch captures clusters of malignant cells with increasing atypia, while the regression branch identifies the surrounding reactive stroma. In TRG 3, the tumor branch

strongly activates on regions with marked nuclear atypia and disordered arrangement, whereas the regression branch focuses on extensive fibrosis and necrosis. This complementary visualization confirms that the model effectively captures the reciprocal progression between tumor residues and regression traces.

4 Conclusion

In this paper, we propose a novel MIL framework that synergizes feature disentanglement with ordinal constraints for TRG assessment. The DBGA disentangles residual tumor patterns from regression traces to model their reciprocal interplay, while the OAR regularizes the feature space to preserve the intrinsic ordinal structure of TRG. This pathology-based design improves grading reliability and supports clinical decision-making.

Acknowledgements. This work was supported by the National Natural Science Foundation of China (Grant No. 62371409), National High Level Hospital Clinical Research Funding (2025-PUMCH-D-002).

References

[1] Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R.L., Soerjomataram, I., Jemal, A.: Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians **74**(3), 229–263 (2024)

[2] Yang, G., Su, X., Yang, H., Luo, G., Gao, C., Zheng, Y., Xie, W., Huang, M., Bei, T., Bai, Y., *et al.*: Neoadjuvant programmed death-1 blockade plus chemotherapy in locally advanced esophageal squamous cell carcinoma. Annals of translational medicine **9**(15), 1254 (2021)

[3] Bolger, J.C., Donohoe, C.L., Lowery, M., Reynolds, J.V.: Advances in the curative management of oesophageal cancer. British Journal of Cancer (2022)

[4] Lerttanatum, N., Tharavej, C., Chongpison, Y., Sanpavat, A.: Comparison of tumor regression grading system in locally advanced esophageal squamous cell carcinoma after preoperative radio-chemotherapy to determine the most accurate system predicting prognosis. Journal of Gastrointestinal Oncology **10** (2019)

[5] Qin, J., Xue, L., Hao, A., Guo, X., Jiang, T., Ni, Y., Liu, S., Chen, Y., Jiang, H., Zhang, C., *et al.*: Neoadjuvant chemotherapy with or without camrelizumab in resectable esophageal squamous cell carcinoma: the randomized phase 3 escort-neo/ncces01 trial. Nature medicine **30**(9), 2549–2557 (2024)

[6] Natsuru, S., Ichikawa, H., Yusuke, M., Takaaki, H., Yosuke, K., Ishikawa, T., Yuki, H., Kohei, M., Yoshifumi, S., Masayuki, N., *et al.*: Clinical utility of yptnm stage grouping in the 8th edition of the american joint committee on cancer

tnm staging system for esophageal squamous cell carcinoma. Annals of Surgical Oncology **28**(2), 650–660 (2021)

[7] Xu, L., Wei, X.-f., Li, C.-j., Yang, Z.-y., Yu, Y.-k., Li, H.-m., Xie, H.-n., Yang, Y.-f., Jing, W.-w., Wang, Z., *et al.*: Pathologic responses and surgical outcomes after neoadjuvant immunochemotherapy versus neoadjuvant chemoradiotherapy in patients with locally advanced esophageal squamous cell carcinoma. Frontiers in Immunology **13**, 1052542 (2022)

[8] Elforaici, M.E.A., Azzi, F., Trudel, D., Nguyen, B., Montagnon, E., Tang, A., Turcotte, S., Kadoury, S.: Cell-level gnn-based prediction of tumor regression grade in colorectal liver metastases from histopathology images. In: 2024 IEEE International Symposium on Biomedical Imaging (ISBI), pp. 1–5 (2024). IEEE

[9] Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International Conference on Machine Learning, pp. 2127–2136 (2018)

[10] Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature biomedical engineering **5**(6), 555–570 (2021)

[11] Tang, W., Huang, S., Zhang, X., Zhou, F., Zhang, Y., Liu, B.: Multiple instance learning framework with masked hard instance mining for whole slide image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4078–4087 (2023)

[12] Liu, P., Ji, L., Zhang, X., Ye, F.: Pseudo-bag mixup augmentation for multiple instance learning-based whole slide image classification. IEEE Transactions on Medical Imaging **43**(5), 1841–1852 (2024)

[13] Yao, J., Zhu, X., Jonnagaddala, J., Hawkins, N., Huang, J.: Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. Medical image analysis **65**, 101789 (2020)

[14] Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al.: Transmil: Transformer based correlated multiple instance learning for whole slide image classification. Advances in neural information processing systems (2021)

[15] Chen, R.J., Chen, C., Li, Y., Chen, T.Y., Trister, A.D., Krishnan, R.G., Mahmood, F.: Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16144–16155 (2022)

[16] Feng, J., Zhou, Z.-H.: Deep miml network. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 31 (2017)

[17] Xiang, J., Zhang, J.: Exploring low-rank property in multiple instance learning for whole slide image classification. In: The Eleventh International Conference on Learning Representations (2023)

[18] Zhang, Y., Li, H., Sun, Y., Zheng, S., Zhu, C., Yang, L.: Attention-challenging multiple instance learning for whole slide image classification. In: European Conference on Computer Vision, pp. 125–143 (2024). Springer

[19] Zhang, Y., Li, H., Sun, Y., Shui, Z., Li, J., Zhu, C., Yang, L.: Aem: attention entropy maximization for multiple instance learning based whole slide image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 45–55 (2025)

[20] Bioptimus: H-optimus-1. https://huggingface.co/bioptimus/H-optimus-1

[21] Vorontsov, E., Bozkurt, A., Casson, A., Shaikovski, G., Zelechowski, M., Severson, K., Zimmermann, E., Hall, J., Tenenholtz, N., Fusi, N., *et al.*: A foundation model for clinical-grade computational pathology and rare cancers detection. Nature medicine **30**(10), 2924–2935 (2024)